BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e135092
doi: 10.3897/biss.8.135092

OPEN ACCESS

Conference Abstract

Graphical User Interface for Biodiversity Digital Twins: Data Challenges

Tomáš Martinovič[‡], Kata Sara-aho[§], Ondrej Salamon[‡], Simon Rolph[|], Allan T Souza[¶]

‡ VSB - Technical University of Ostrava, Ostrava, Czech Republic
§ CSC - IT Center for Science, Espoo, Finland
| UK Centre for Ecology & Hydrology, Wallingford, United Kingdom
¶ Institute for Atmospheric and Earth System Research INAR, Forest Sciences, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland

Corresponding author: Tomáš Martinovič (tomas.martinovic@vsb.cz)

Received: 19 Aug 2024 | Published: 20 Aug 2024

Citation: Martinovič T, Sara-aho K, Salamon O, Rolph S, Souza A (2024) Graphical User Interface for Biodiversity Digital Twins: Data Challenges. Biodiversity Information Science and Standards 8: e135092. https://doi.org/10.3897/biss.8.135092

Abstract

Digital Twins are a new concept in the field of biodiversity (de Koning et al. 2023), one aspect is the user interface for interacting with digital twins. In the BioDT (Biodiversity Digital Twin) (BioDT 2022) project we are creating a graphical web interface (Martinovič et al. 2024) using the R Shiny framework, which allows small-scale data analysis to be done directly on the server running the web interface, while making it possible to offload a large-scale analysis to a supercomputer such as LUMI (Large Unified Modern Infrastructure). Additionally, we foresee that multiple prototype Digital Twins (pDTs) will be available as part of one web application. This brings multiple challenges in optimization of data flows for the computation and interaction with users, especially since data used by the pDTs, developed in the BioDT, are usually stored across multiple data sources.

In the BioDT web application, we want to enable users to access different pDTs concerned with questions such as ecosystem services, biodiversity dynamics, DNA-related biodiversity tasks, pollinators, invasive species, disease outbreaks, and more. While some of these pDTs compute results at a remote server and provide only the newest results to the application, others aim to allow users to execute their own pDT runs with their own data and settings. This leads to many different user scenarios and impacts authentication and authorization flows, as well as data flows. In addition to this, we want

to make the whole system comply with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles as much as possible. For this, we need to support traceability of data, models, and pDT executions.

This means that we advocate for data to be hosted at data providers with APIs for machine-to-machine interaction and extensive metadata support; models to be as open as possible and versioned; and to have a workflow execution orchestrator that can keep track of the models' execution and their related inputs and outputs. In terms of data, we are using research infrastructures (RIs) such as GBIF (Global Biodiversity Information Facility) and eLTER (Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research) for the input data streams and when it is not possible to use established RIs, we are using self-hosted services. This work is still in progress and some of the self-hosted datasets are still being used, but in the future, we are looking into the possibility of having a dynamic system that will support data formats that cannot be currently stored at the established RIs. Our colleagues at BioDT consortium are preparing a special data server (El-Gabbas et al. 2023) that can handle complex data processing and serve data through an API without the need to download data for multiple commonly used formats.

Some of the BioDT pDTs leverage the computational strength of High-Performance Computing (HPC) clusters and in such cases, classical cloud workflow orchestrators are not an option due to the specific security policies of such centers. To solve this, we turned to the LEXIS Platform, which can execute predefined workflows on a combination of cloud and HPC resources, and track the executions and related execution metadata. We are looking into exporting the descriptions of the workflow executions to Research Object Crates (RO-Crates) and then uploading this information to a remote server, where users could check their execution settings later.

A current main development focus is the question of how to tackle the challenge of multiple authentication systems. This is specifically of concern in the case of sensitive data, which need to remain secure and available only to selected people. Due to interactions with several systems, we are encountering authentication issues of multiple different identities that should be recognized as one. The simplest solution here is to use one platform for data storage, workflow execution and web application. However, in the future we hope to find a more general solution that would not require data transfer to a single platform, since this could lower the usage of the BioDT platform due to legal restriction on some data.

Another challenge in optimization of the dataflows is to avoid downloading the same data repeatedly and to be able to provide users with relevant data in the web application as fast as possible, since long waiting time would result in people not using the web application. For these issues, we are considering a smart-caching mechanism, however such functionality is not yet defined.

Keywords

FAIR data, modelling, R, RO-Crate, shiny, web application, workflows

Presenting author

Tomáš Martinovič

Presented at

SPNHC-TDWG 2024

Funding program

This study has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057437 (BioDT project). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Development Access call. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 871128 (eLTER PLUS).

Conflicts of interest

The authors have declared that no competing interests exist.

References

- BioDT (2022) Biodiversity Digital Twin for Advanced Modelling, Simulation and Prediction Capabilities. https://doi.org/10.3030/101057437
- de Koning K, Broekhuijsen J, Kühn I, Ovaskainen O, Taubert F, Endresen D, Schigel D, Grimm V (2023) Digital twins: dynamic model-data fusion for ecology. Trends in Ecology & Evolution 38 (10): 916-926. https://doi.org/10.1016/j.tree.2023.04.010
- El-Gabbas A, Khan T, Golivets M, Kühn I (2023) Invasive alien species Digital Twin (IAS-DT). Zenodo https://doi.org/10.5281/zenodo.8100291
- Martinovič T, Sara-aho K, Salamon O, Rolph S, Souza A (2024) https://app.biodt.eu